### Title

Predicting gene expression in the human malaria parasite Plasmodium falciparum using histone modification, nucleosome positioning, and 3D localization features.




### Journal

PLoS computational biology, 15(9)

### ISSN

1553-734X



### Authors

Read, David F
Cook, Kate
Lu, Yang Y
et al.



### Publication Date

2019-09-01

### DOI

10.1371/journal.pcbi.1007329

Peer reviewed

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

# Predicting gene expression in the human malaria parasite *Plasmodium falciparum* using histone modification, nucleosome positioning, and 3D localization features

**David F. Read[1], Kate Cook[1], Yang Y. Lu[1], Karine G. Le Roch[2]*, William Stafford Noble[1]***

**1** Department of Genome Sciences, University of Washington, Seattle, Washington, United States of America, **2** Department of Molecular, Cell and Systems Biology, University of California, Riverside, California, United States of America

* karinel@ucr.edu (KGL); william-noble@uw.edu (WSN)



OPEN ACCESS

**Citation:** Read DF, Cook K, Lu YY, Le Roch KG, Noble WS (2019) Predicting gene expression in the human malaria parasite *Plasmodium falciparum* using histone modification, nucleosome positioning, and 3D localization features. PLoS Comput Biol 15(9): e1007329. https://doi.org/10.1371/journal.pcbi.1007329

**Editor:** Isidore Rigoutsos, Thomas Jefferson University, UNITED STATES

**Received:** February 11, 2019

**Accepted:** August 12, 2019

**Published:** September 11, 2019



**Data Availability Statement:** All relevant data are available from URLs provided by the references in Table 2.

**Funding:** Funding is from the National Institute of Allergy and Infectious Diseases (R01AI106775 and R01AI13651) and the National Institute of General Medical Sciences (P41GM103533). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Abstract

Empirical evidence suggests that the malaria parasite *Plasmodium falciparum* employs a broad range of mechanisms to regulate gene transcription throughout the organism's complex life cycle. To better understand this regulatory machinery, we assembled a rich collection of genomic and epigenomic data sets, including information about transcription factor (TF) binding motifs, patterns of covalent histone modifications, nucleosome occupancy, GC content, and global 3D genome architecture. We used these data to train machine learning models to discriminate between high-expression and low-expression genes, focusing on three distinct stages of the red blood cell phase of the *Plasmodium* life cycle. Our results highlight the importance of histone modifications and 3D chromatin architecture in *Plasmodium* transcriptional regulation and suggest that AP2 transcription factors may play a limited regulatory role, perhaps operating in conjunction with epigenetic factors.

## Author summary

The parasite responsible for the most lethal form of malaria, *Plasmodium falciparum*, employs a variety of mechanisms to modify the expression of its genes throughout its complex life cycle. In this work, we gather a rich collection of data describing various aspects of the gene regulatory apparatus in *P. falciparum*, and we use a machine learning approach to help understand the relative importance of each potential regulatory mechanism. Our results highlight the importance of two particular mechanisms: patterns of biochemical modifications on the histone proteins that form the primary scaffold for DNA in the cell and the three-dimensional conformation of DNA in the nucleus.

**Competing interests:** The authors have declared that no competing interests exist.

## Introduction

*Plasmodium falciparum* is the deadliest species of malaria parasite, responsible for 445,000 deaths in 2016 [1]. As resistance to antimalarial drugs spreads, demand for novel antimalarials increases. Designing such novel drugs would require an improved understanding of the biology of this parasite. Currently, one of the primary open questions in *Plasmodium* biology is how the parasite maintains precise control of gene expression. The current work aims to address this question by constructing an accurate predictive model of *P. falciparum* transcription. The model accounts for the rich landscape of transcriptional control mechanisms in *Plasmodium* by incorporating five types of features, representing transcription factor (TF) binding, covalent histone modifications, nucleosome positioning, GC content, and chromatin 3D structure.

In many eukaryotes, TF binding within and around gene promoters is considered the dominant mechanism of gene expression control. However, in *Plasmodium*, several lines of evidence suggest that TF binding may be less central to transcriptional control. First, although major components of the general transcription machinery are present in the *Plasmodium* genome [2], a relatively small set of specific *Plasmodium* TFs ($\sim$ 27) have been identified and validated in the parasite genome [2]. In comparison, the similarly sized genome of the yeast *S. cerevisiae* contains $\sim$170 specific TFs [3]. Second, among the subset of TFs whose binding affinities have been characterized via *in vitro* protein binding microarrays [4], only a handful display stage-restricted expression and play clear roles in regulating life cycle transitions. An example is PfAP2-G, which drives expression of gametocyte-specific genes [5, 6]. Third, a large number of *Plasmodium* genes are predicted by homology to function in the regulation of chromatin structure, mRNA decay, and translation [2], suggesting the importance of epigenetic and post-transcriptional regulation.

Among mechanisms for epigenetic regulation, patterns of covalent histone modifications are perhaps the most widely studied and understood. In this respect, some aspects of *P. falciparum* gene regulation are shared with other eukaryotes, including the presence of the typically heterochromatin-associated H3K9me3 histone modification at repressed *var* genes (referred to as virulence genes, for their role in parasite pathogenicity) [7] and depletion of promoter nucleosomes correlating with gene transcription [8]. On the other hand, *Plasmodium* epigenomic dynamics also exhibit notable deviations from those in commonly studied eukaryotes, such as abundant and broad distributions of activating histone marks [9, 10] or active histone variant H2Z in the promoters of all genes with the exception of genes involved in immune evasion [11], an absence of H3K27me3 repressive marks [9], and genome-wide changes in nucleosome occupancy during the asexual cycle [8, 12]. These observations suggest that the parasite may make use of a "histone code" like other well-characterized eukaryotes, though the specific role of individual elements may differ.

In addition, empirical evidence suggests that gene regulation in *Plasmodium* occurs through changes in chromatin structure, including shifts in nucleosome occupancy at the local level and 3D positioning at larger scales. Nucleosome occupancy, as measured by MNase-assisted isolation of nucleolar elements (MAINE) and formaldehyde-assisted isolation of regulatory elements (FAIRE), or assay for transposase accessible chromatin with high-throughput sequencing (ATAC-seq), exhibit cyclic patterns that closely track changes in gene expression during the red blood cell (erythryocytic) stages of the parasite life cycle [12, 13]. In addition, 3D models of *Plasmodium* DNA based on Hi-C assays at multiple time points during the red blood cell [9] and transmission [14] stages of the parasite point to a "gradient" of expression across the nucleus, from a repressive center near the telomeres to an expressive center at the centromeres.

**Table 1. Comparison of methods for predicting gene expression.** The "mRNA Expression" feature involves using the mRNA expression of a set of putative regulators to predict the mRNA expression of a different set of target genes.

| Model | Year | Yeast | Mouse | Human | mRNA expression | TF motifs | DNA sequence | TF ChIP-seq | Histone ChIP-seq | DNase or ATAC | PolII ChIP-seq | Classification | Regression |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Multiple linear regression [18] | 2001 | ✓ | | | | ✓ | ✓ | | | | | | ✓ |
| Conditional probability given levels of regulators [19] | 2002 | ✓ | | | ✓ | | | | | | | ✓ | |
| Classification tree [20] | 2003 | ✓ | | | ✓ | | | | | | | ✓ | |
| Bayesian network [21] | 2004 | ✓ | | | ✓ | | ✓ | | | | | ✓ | |
| Boosted alternating decision trees [22] | 2004 | ✓ | | | ✓ | ✓ | | | | | | ✓ | |
| Boosted alternating decision trees [23] | 2006 | ✓ | | | ✓ | ✓ | | | | | | ✓ | |
| Principal component regression and regression tree [24] | 2009 | | ✓ | | | | | ✓ | | | | | ✓ |
| Multiple linear regression [25] | 2010 | | | ✓ | | | | | ✓ | | | | ✓ |
| Support vector machine [15] | 2011 | | ✓ | | | | | | ✓ | | | ✓ | ✓ |
| Random forest and multiple linear regression [16] | 2012 | | | ✓ | | | | | ✓ | | | ✓ | ✓ |
| Multiple log-linear regression [26] | 2012 | | ✓ | | | | | ✓ | ✓ | ✓ | | | ✓ |
| Bayesian variable selection regression [27] | 2014 | | | ✓ | | | | | ✓ | | ✓ | | ✓ |
| Multiple regression [28] | 2015 | | | ✓ | | ✓ | ✓ | | ✓ | ✓ | | | ✓ |
| Multilayer perceptron [29] | 2016 | | | ✓ | ✓ | | | | | | | | ✓ |
| Convolutional neural network [17] | 2016 | | | ✓ | | | | ✓ | | | | ✓ | |
| Multiple regression [30] | 2017 | | | ✓ | ✓ | | | | ✓ | ✓ | | | ✓ |
| Convolutional neural network [31] | 2018 | | | ✓ | | | ✓ | | | | | | ✓ |
| Multitask regression [32] | 2018 | | | ✓ | | ✓ | ✓ | | | ✓ | | | ✓ |

https://doi.org/10.1371/journal.pcbi.1007329.t001

Based on the above evidence, we hypothesized that the cascade of transcripts observed throughout the red blood cell (erythrocytic) cycle is the result of a combination of transcription factor binding, histone modifications, and changes in chromatin structure. We further predicted than an integrated analysis of the relationships among transcription and TF binding, histone modification, and chromatin structure data could reveal the relative significance of individual features in defining high- and low-expression genes.

We are not the first to build predictive models of gene expression (Table 1), though to our knowledge we are the first to do so in *Plasmodium*. To keep the model simple, we focus on the binary classification task, in which each gene is either "on" or "off," rather than the more challenging regression setting. Prediction of gene expression has been framed as a classification

task in numerous previous works, with our approach most closely resembling the analysis in references [15–17]. Furthermore, because we sought to determine the importance of features within individual stages we restrict ourselves to predicting relative high- or low-expression labels with respect a single parasitic stage at a time, rather than developing a single model that predicts absolute expression values irrespective of stage. Accordingly, we build separate models in three different stages of the *P. falciparum* life cycle and analyze the resulting models to understand which features are implicated in the up- or down-expression of *Plasmodium* genes in different stages of the parasitic life cycle.

## Methods

A description of all methods is given below. All processed data, as well as code used for data processing and model training/evaluation, is available online (https://github.com/Daread/plasmodiumExprPrediction).

### Data sets

Although *P. falciparum* passes through multiple stages—mosquito, human liver, and human blood—we focus here on the human blood stage of the parasite life cycle, primarily because of the availability of a wide number of relevant data sets. We gathered data for three time points, corresponding to the three main asexual stages within the red blood cell cycle: ring, trophozoite and schizont. Most data sets described below (Table 2) are available in all three of these time points, with the exception of some ChIP-seq data for covalent histone modifications (H3K36me2, H3K36me3, H3K9me3, H4K20me3, and one replicate of H3K4me3 [33]) that were not available for the trophozoite stage.

**GRO-seq.** To define the on/off labels for our classifier, we used the GRO-seq values from Lu *et al.* [37]. We used GRO-seq values normalized for GC content, gene length, and the "parasitemia factor" of a stage [37], available in S1 Table. To generate binary labels for genes, for each stage we sorted all protein-coding genes by the normalized GRO-seq counts assigned to that gene in that stage. We labeled the top third of genes as "High expression" and the bottom

**Table 2. Summary of datasets used in classification models.** Dataset sources are shown, along with the time points from each study which were used to represent each of three life cycle stages. "N/A" is listed for the GC content and motif score features, as they did not vary across life cycle stages.

| Feature | Study | Description | Time point | | |
|---|---|---|---|---|---|
| | | | Ring | Trophozoite | Schizont |
| Distance to telomeres | [34] | Hi-C inferred distance | ✓ | ✓ | ✓ |
| Distance to centromeres | [34] | Hi-C inferred distance | ✓ | ✓ | ✓ |
| Distance to center | [34] | Hi-C inferred distance | ✓ | ✓ | ✓ |
| Nucleosome occupancy | [8] | 100-200 base pair fragments | ✓ | ✓ | ✓ |
| H2A.z | [35] | ChIP-Seq | ✓ | ✓ | ✓ |
| H3K9Ac | [35] | ChIP-Seq | ✓ | ✓ | ✓ |
| H3K4me3 (Bartfai et al) | [35] | ChIP-Seq | ✓ | ✓ | ✓ |
| H3K36me2 | [33] | ChIP-Seq | ✓ | | ✓ |
| H3K36me3 | [33] | ChIP-Seq | ✓ | | ✓ |
| H3K9me3 | [33] | ChIP-Seq | ✓ | | ✓ |
| H4K20me3 | [33] | ChIP-Seq | ✓ | | ✓ |
| H3K4me3 (Jiang et al) | [33] | ChIP-Seq | ✓ | | ✓ |
| GC content | See Methods | 100 bp sliding windows | N/A | N/A | N/A |
| TF motifs | [36] | FIMO scan of TF motifs | N/A | N/A | N/A |

https://doi.org/10.1371/journal.pcbi.1007329.t002

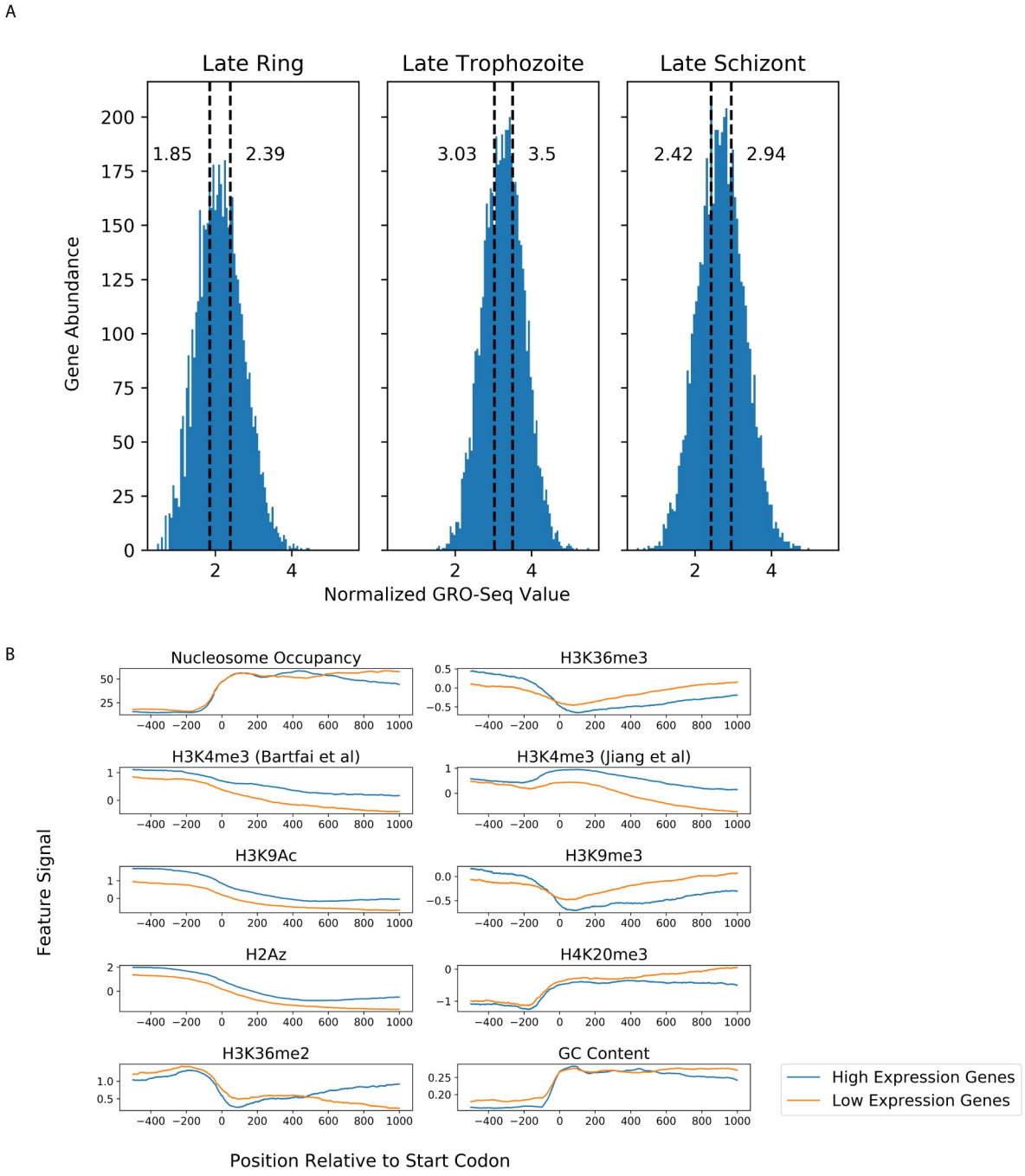


**Fig 1. Differences between high- and low-expression genes.** (A) Histograms showing the distribution of normalized GRO-Seq values assigned to protein-coding genes within each of three stages of the *P. falciparum* life cycle. The labeled, dashed vertical lines indicate the cut-off values for genes categorized as low- and high-expression. (B) Aggregation plots showing the average signal for features with respect to the start codons of high-expression (blue) and low-expression (orange) genes in the ring stage.

https://doi.org/10.1371/journal.pcbi.1007329.g001

third of genes as "Low Expression" (Fig 1). The middle third of genes were not used in the analysis.

The decision to use tertiles rather than, say, quartiles or simply dividing at the median was somewhat arbitrary. Past works have used a number of schemes, such as dividing genes into

tertiles [22, 23], dividing genes in half at the median [15, 17], or dividing by zero/non-zero status [16]. We elected to use tertiles and perform classification using the top and bottom sets in part to make the classification task somewhat easier (by only giving the model examples that are well-separated by expression) and in part to limit the detrimental effect of possible noise in the GRO-seq data (because noise is less likely to flip a gene between classes when the divisions are made at tertiles than if divisions were made at the median).

**Transcription start site and coding sequence annotations.** Many of the features that we employed require specifying the start coordinate of each given gene. For this purpose, we use two data sets of coordinates: either the start codons from the PlasmoDB v29 annotation or transcription start sites based on CAGE-Seq data from [38]. In that resource, multiple start sites are often annotated for a given protein coding gene. To assign a single TSS for use in feature assignment, we first looked to see if the "primary TSS" assigned in [38] was upstream of the start codon of a gene. If it was, then the start of that TSS was used. Otherwise, we used the TSS lying upstream and closest to the start codon. If no annotated TSS was upstream of the start codon of a gene, then the start codon was used.

**Histone modification ChIP-seq.** ChIP-seq data for the following histone modifications were collected from two studies: H3K4me3, H3K9Ac, and H2.Az from Bartfai et al. [35] and H3K36me2, H3K36me3, H3K9me3, H4K20me3, and H3K4me3 from Jiang et al. [33] Note that one mark, H3K4me3, was measured in both studies. All of the ChIP-seq data was reanalyzed using a standard pipeline that consisted of trimming reads to 76 nucleotides using the fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html), mapping reads to the *P. falciparum* genome (PlasmoDB v29) using bwa-mem [39], filtering unmapped and multimapping reads using samtools [40], and generating bedgraph files using bedtools [41]. Fold-change over background was calculated relative to input DNA where available. For histone ChIP-seq datasets from Jiang *et al*., no input DNA data was available, so values were calculated relative to the mean signal over all the data.

**Nucleosome occupancy.** MNase data from [8] was downloaded in FASTQ format, then trimmed and filtered using sickle version 1.33 (https://github.com/najoshi/sickle). Reads were aligned to the *P. falciparum* genome (PlasmoDB v29) using bwa-mem [39], then sorted and filtered for mapped reads using samtools [40]. A custom Python script selected all alignments between 100 and 200 bp in length, which were then used to generate a bedgraph file using genomeCoverageBed [41].

**Hi-C.** Per-gene distances from telomere centroid, centromere centroid, and center were computed in a previous study carried out by our lab [34], using 3D models generated from Hi-C data using PASTIS [42]. These values were obtained directly from https://noble.gs.washington.edu/proj/plasmo3d/.

**DNA sequence features.** Position-frequency matrices for 25 AP2 family transcription factors for *P. falciparum* were downloaded from CIS-BP [36]. We focused on available AP2 family motifs for two reasons. First, AP2 transcription factors have been widely speculated to play a key role in TF-mediated transcriptional regulation throughout the erythrocytic cycle due to variable expression, sequence-specific DNA binding, and presence of AP2 motifs upstream of genes whose expression varies throughout erythrocytic stages [43, 44]. Second, the DNA binding specificities of AP2 transcription factors were evaluated en masse using a high-throughput *in-vitro* protein binding microarray and subjected to *in vivo* validation [43], generating a motif set derived by a consistent, rigorous workflow.

With each motif, we scanned the *P. falciparum* genome using FIMO [45] with a p-value threshold of 0.01. This fairly permissive cut-off was arbitrarily set, leaving more aggressive feature selection for downstream model training and evaluation. To ensure that the background model represented the unique sequence context of *P. falciparum*, we generated a background

model from the *P. falciparum* genome with the MEME Suite command `fasta-get-markov` with Markov order 1 [46]. In addition, percent GC was calculated in 101 base windows centered at each position in the genome.

Features based on histone modifications, H2Az composition, nucleosome occupancy, and GC content were segregated into "promoter" and "gene body" features. The "promoter" feature was the mean feature signal from -500 bases up to the start codon, whereas the "gene body" feature was the mean feature signal from the start codon to 1 kb into the coding sequence.

## Predictive models

**Models.** We used three types of models to classify *Plasmodium* expression and select predictive features. The first was logistic regression with elastic net regularization, using the scikit-learn implementation (sklearn.linear_model.SGDClassifier). The second was a tree model with gradient boosting, using the XGboost Python implementation (xgboost.XGBClassifier). The third was a multi-layer perceptron model, with two hidden layers, each containing the same number of nodes as the input layer. This model was implemented by DeepPINK [47], which is designed to achieve robust feature selection with a controlled error rate.

**Performance metric.** The performance of each model was evaluated in terms of receiver operator characteristic (ROC) curves. These plots show the rate of true positive classifications (on the y-axis, indicating sensitivity) against the rate of false positive classifications (on the x-axis, indicating 1—specificity). The area under the ROC curve (AUROC) quantifies the ability of the classifier to balance sensitivity (true positives) against specificity (avoiding false positives). An AUC value of 1 corresponds to perfect performance, whereas a value of 0.5 corresponds to random guessing.

For Fig 2A, the ROC curve for logistic regression classification was generated by combining the gene scores from the test sets in three separate folds of cross-validation. These scores were sorted together to generate the combined ROC curve shown.

**Train/test data splitting.** We split the *P. falciparum* into five approximately equally sized (by gene count) folds by chromosome: fold 1 included chromosomes 1, 3, and 13; fold two included 2, 9, and 11; fold three included 7 and 14; fold four included 6, 8, and 10; and fold five included 4, 5, and 12. This split was done by calculating the number of genes-per-fold in a perfectly even split, then choosing the division of chromosomes whose totals had the smallest mean-squared error from this ideal value, across all possible permutations of chromosome-to-fold assignments. A Python script that tests all permutations of chromosome sets, selecting the division that minimizes the mean squared error, is available in the Github repository, dataPreProcessing/selectingDataFolds/divideGenomeIntoFiveSets.py.

## Model development and hyperparameter tuning

The first three folds were used for feature development and hyperparameter tuning. During this stage, we selected hyperparameters by three-fold internal cross-validation. For the logistic regression model with elastic net regularization, we tuned the "alpha" and "l1_ratio" parameters in a sklearn.linear_model.SGDClassifier model. The "alpha" value determines the magnitude of the regularization penalty relative to classification error, while "l1_ratio" determines the relative magnitude of the L1 and L2 penalty terms (1 = pure LASSO penalty, 0 = purely ridge regularization). For the boosted trees model, we tuned the "max_depth", "min_child_weight", "subsample", "colsample_bylevel", and "n_estimators" hyperparameters in an xgboost.XGBClassifier model. "Max_depth" controls the tree depth of the decision trees composing the XGBoost ensemble, "min_child_weight" controls the minimum weight in a leaf node that is

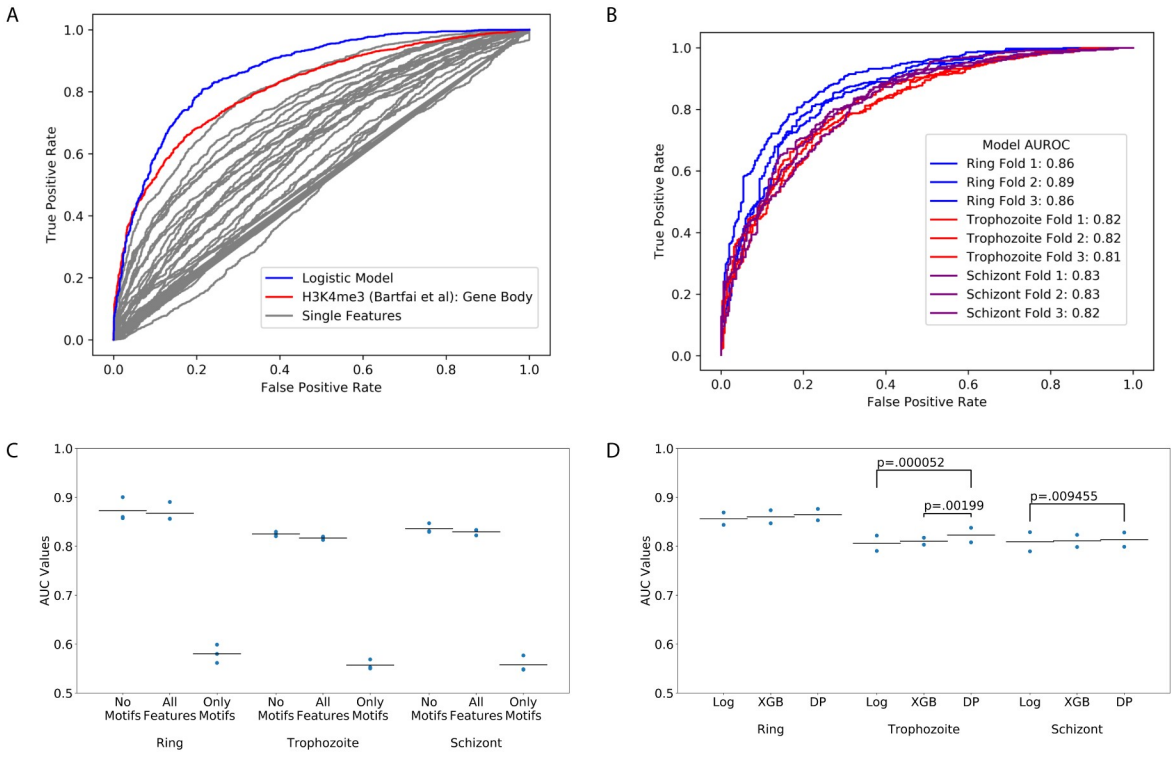


**Fig 2. Comparison of classification models.** (A) Classification models outperform all individual features for use in classification of gene expression. Grey lines represent the ROC curves resulting from ranking genes by the values of single features in the ring stage, with the best-performing feature shown in red. The blue line represents the ROC curve from training a logistic regression model with elastic net regularization, where the curve is created by combining the predictions across all three test sets. (B) The ROC curves for classification of gene expression by logistic regression across ring, trophozoite, and schizont stages. Individual curves represent performance in one of test three folds in cross-validation. (C) AUROC values for logistic models trained with or without motif scores as features. Points represent AUROC values on the test set in three-fold cross-validation; bars represent average AUROC values on the test data. (D) The AUROC values resulting from training of distinct models in different stages ("Log" = Logistic Regression, "XGB" = XGBoost, "DP" = DeepPINK). Individual points represent the AUROC values from distinct test sets, for the listed model in a given stage. Brackets are labeled with p values for pairwise comparisons within stages where $p < 0.05$, using the DeLong method for comparing AUC values.

https://doi.org/10.1371/journal.pcbi.1007329.g002

allowed to be split further, "subsample" controls the portion of training data samples for training each additional tree, "colsample_by_level" controls whether re-sampling is done for each new depth level within trees, and "n_estimators" is the number of trees in the model. In each case, we performed a grid search across the values listed in Table 3, testing all possible

**Table 3. Hyperparameter selection.**

| Model | Parameter | Possible values | Selected value | | |
|---|---|---|---|---|---|
| | | | Ring | Trophozoite | Schizont |
| LR | alpha | 0.1, 0.01, 0.001, 0.0001 | 0.0001 | 0.0001 | 0.0001 |
| LR | l1_ratio | 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95 | 0.95 | 0.8 | 0.9 |
| Boosted trees | max_depth | 4, 5, 6, 7, 8, 9 | 4 | 6 | 4 |
| Boosted trees | min_child_weight | 2, 3, 4, 5, 6, 7 | 5 | 5 | 6 |
| Boosted trees | subsample | 0.3, 0.4, 0.7, 0.8, 1.0 | 0.7 | 0.4 | 0.4 |
| Boosted trees | colsample_bylevel | 0.3, 0.5, 0.7, 1.0 | 0.3 | 0.5 | 0.3 |
| Boosted trees | n_estimators | 40, 60, 80, 100 | 100 | 60 | 80 |

https://doi.org/10.1371/journal.pcbi.1007329.t003

combinations of hyperparameter values using cross-validation within the three folds used for model development and selecting the hyperparameter combination with the lowest test error.

On the basis of initial analyses, we eliminated the motif-based features from our feature set, and we chose to use features based on CDS rather than TSS locations (see Section for details).

## Model evaluation on test data

Subsequently, we trained classifiers to make predictions on each of the two test folds, in each case training on the four remaining folds. In this case, hyperparameters were selected that yielded the greatest AUROC value in the three training set "sub-folds" using cross-validation, as implemented in GridSearchCV in sklearn.grid_search. The selected hyperparameters are listed in Table 3. Probabilistic classification scores for all genes in both of the two test folds were combined for testing the statistical significance of differences in AUC values. AUC values were compared using the DeLong test for correlated AUCs [48] as implemented in the pROC package in the R language [49].

**XGBoost and SHAP values.** The gradient boosting method XGBoost is powerful but challenging to interpret. XGBoost assigns classification labels by taking a consensus decision from an ensemble of individual decision trees [50]. XGBoost models are appealing due to their ability to capture complex interactions among features as well as non-linear relationships between features and classification labels [50]. However, understanding the importance of individual features within such ensembles is challenging, because the model may use a given feature in multiple locations across the individual trees (in contrast to a regression model with a readily interpretable coefficient assigned to a feature).

Consequently, we used SHAP [51] to help interpret the trained XGBoost models. SHAP is a software package that quantifies the effect of each feature on the classification of each example (each gene, in our case) by measuring how much information that feature provides in addition to various subsets of other features being used in the model. The method obeys key mathematical properties and matches human intuition in tested cases [51]. Running SHAP on our trained XGBoost models provided us with "SHAP values" for each feature, for every gene. These scores can be studied on a gene-by-gene basis and can be aggregated across all genes within a stage to obtain a consensus score, comparable to a regression coefficient.

**DeepPINK.** Similar to XGBoost, DeepPINK can also capture non-linear relationships between features and classification labels. Rather than boosted gradients, DeepPINK uses a deep neural network model. Importantly, DeepPINK is able to reliably choose relevant features with a controlled error rate, regardless of arbitrarily complex interactions among features. To rigorously control the false discover rate among selected features, DeepPINK relies upon the recently described model-X knockoffs framework [52]. The primary methodological novelty in DeepPINK is its deep neural network architecture, which enables application of the model-X framework.

## Determining feature importance

After training, we examined each model to extract information about which features the model deemed most relevant to the given classification task. For the logistic regression models, we recorded the coefficients assigned to each feature. For XGboost, we used the SHAP package to calculate "SHAP values" for each feature at each gene [51]. The magnitude of the feature importance score was defined as the mean SHAP value across all genes. The sign for the feature importance score (indicating whether a feature indicates high- or low-expression) was defined by the direction of correlation between feature values and SHAP values across all genes. DeepPINK computes feature weights by multiplying the weight matrices across all

layers of the deep neural network. The resulting weights can be either positive or negative, indicating the direction of correlation between features and the label. We used the the squared value of the feature weight as the feature importance score.

## Results

### High- and low-expression genes display qualitative genetic and epigenetic differences

Drawing from a variety of data sources, as described in Methods, we constructed a data set of heterogeneous gene features across each of three stages of the erythrocytic cycle (ring, trophozoite and schizont). GRO-seq measurements of nascent transcription were used to identify genes which high expression (top third) and low expression (bottom third) (Fig 1A).

As an initial step of data exploration, features with signal at a base-by-base level (such as ChIP-seq tracks or GC content) were visualized using aggregation plots showing the average level of signal for a feature with respect to the start codon, segregated by high-expression and low-expression genes (Fig 1B). These plots show expected trends, including enrichment of H3K4me3, H3K9Ac, and H2Az in highly expressed genes, as well as depletion of nucleosome occupancy in promoter regions.

Given the apparent differences in signals upstream and downstream of the start codon, we split each of these main features into two features. The "promoter" feature was the mean feature signal from -500 bases up to the start codon, whereas the "gene body" feature was the mean feature signal from the start codon to 1 kb into the coding sequence. This was done for all covalent histone modification features, H2Az composition, nucleosome occupancy, and GC content.

Similarly, for motif features, we calculated the maximum motif match log-odds score in two windows. The "promoter" window was from -500 bases up to the start codon, while the "gene body" region extended from the start codon up to 1000 bases into the coding sequence.

Ultimately, each ring- and schizont-stage gene was characterized by a set of 73 features, including 50 motif-based features, 14 histone modification features, 7 features characterizing local and global chromatin structure, and 2 features describing local GC content. Trophozoite stage genes used the same feature vector, but with 10 histone features removed due to missing ChIP-seq data sets in that stage. These matrices, including feature values and gene labels, are available for all three stages via the Github repository under the modelData directory. For each stage, we include two versions of the file, with and without motif features, for convenience.

### Machine learning models accurately distinguish between expressed and non-expressed genes

We initially examined single features to establish a baseline of classification performance based on a simple ordering of genes by each individual feature. In this way, we generated one ROC curve for each feature (gray lines in Fig 2A), obtaining AUROCs as high as 0.82 for H3K4me3 gene body signal (blue line in Fig 2A) in classification of ring-stage genes.

Next, we compared this baseline approach against a machine learning method that integrates all of the available features. We observed, not surprisingly, that an elastic net-regularized logistic regression ("Logistic") model that integrates all features outperformed rankings based on single features alone: the ROC curve generated by the logistic regression (red curve, Fig 2A) dominates all of the ROC curves generated by ranking genes using single features. We observed similar levels of performance across the three erythrocytic stages, where in a three-

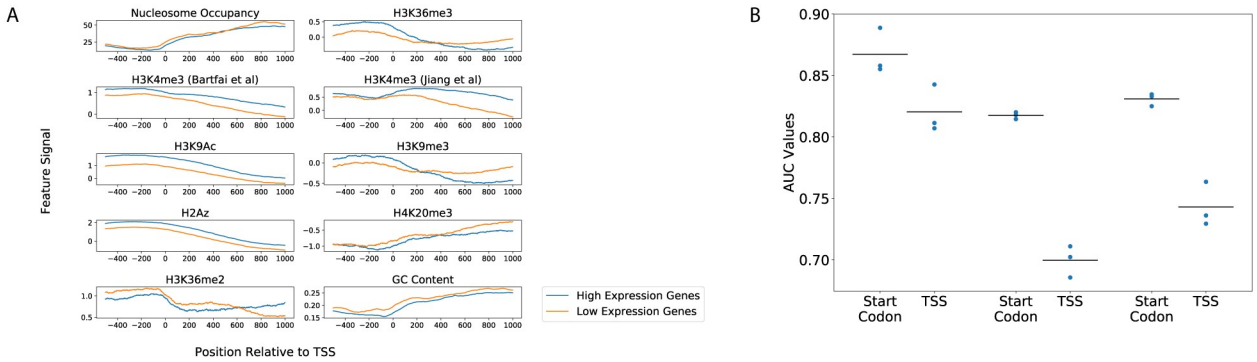


**Fig 3. Transcription start sites versus start codons.** (A) Aggregation plots showing the average signal for features with respect to the transcription start sites of high-expression (blue) and low-expression (orange) genes. (B) A plot of AUROC scores obtained for training a logistic classifier to classify genes as high- or low-expression (in the Ring, Trophozoite, and Schizont stages, left to right). The left column for each stage represents scores using promoter/gene body divided at the start codon (as was done throughout the analysis up to this point), while the right column in each stage used TSSs to divide promoter/gene body regions.

https://doi.org/10.1371/journal.pcbi.1007329.g003

fold cross-validated test, the logistic regression model achieved average AUROCs of 0.868 in ring (Fig 2B), 0.817 in trophozoite and 0.829 in schizont.

## Start codons outperform TSSs for predicting transcription

During our exploratory work using the three "development folds" of data, we tested two different approaches for defining the start of a gene: transcription start sites (TSSs) and start codons. Surprisingly, this testing indicated that start codons are more useful than TSSs for defining the division between promoter and gene body for our predictive models. We started by using genome-wide CAGE-seq datasets to define transcription start sites for all genes (see Methods). Plots of feature scores with respect to these two types of "start" positions—start codons (Fig 1B) and TSSs (Fig 3A)—qualitatively showed stronger trends between high- and low-expression genes when defining promoter/gene body splits using start codons rather than TSSs. Furthermore, when we trained classifiers to label genes using features split by either start codon or TSS, the models using promoter/gene body definitions split by start codons consistently outperformed models using TSSs (Fig 3B). Consequently, we focused analyses in this work on models that are split by start codons rather than TSSs.

## Motif features are not helpful

A key question we aimed to address is the relative utility of the scores derived from TF motifs. Accordingly, we repeated the cross-validated testing of the logistic regression model using three different feature sets: the full set of features, a reduced set in which the TF motif PWM scores have been eliminated, and a set containing only TF motif features. This analysis suggested that the motif features did not aid in classification when combined with non-motif features, and if anything hurt the performance of our models (Fig 2C). Furthermore, models that used only motif features were far less accurate than models that incorporated non-motif features (Fig 2C). In addition, we investigated the possibility that the −500 bp window size used for promoter features may have under-utilized AP2 motifs, if relevant regulatory sequences are spread over larger upstream distances. To this end, we re-trained and evaluated new "motif-only" models, varying the promoter region to include either 1 kb or 2 kb of upstream sequence (instead of 500 bp, as in the original analysis). This analysis (S1 Fig) shows that expanding the range of the upstream window does not improve the performance of models using motif

features alone, suggesting that our modest upstream window size is not missing valuable upstream regulatory AP2 family binding sites.

At this point, we considered our model development complete. Hence, all subsequent analyses incorporate two folds of data that had not been used in prior model development. Thus, whereas previous analyses involve three-fold cross-validation on 3/5 of the data, all subsequent analyses perform two folds of a five-fold cross-validation, training on 4/5 of the data and testing on each of the two held-out test folds (see Section for details).

## Different models have similar accuracies

To determine whether our results thus far depend upon the choice of machine learning method, we also tested two additional types of models: a boosted trees ensemble ("XGBoost"), and a multilayer perceptron with two hidden layers ("DeepPINK"). Refer to "methods" for descriptions of the methods and links to further reading. In each stage, the three models demonstrate similar AUROC performance, with a slight trend of the multi-layer perceptron model outperforming XGboost, which in turn outperforms logistic regression (Fig 2D). We examined all pairwise model comparisons within each stage (DeLong test for correlated AUCs, see Methods), finding three comparisons to have statistically significant differences (Fig 2D). However, even the differences that are statistically significant are relatively modest in absolute terms, leading us to conclude that each of these machine learning methods achieves reasonably good performance in discriminating between *Plasmodium* genes with high and low expression. Accordingly, we used all three methods in subsequent analyses.

## Classification models use stage-specific features

Having established the feasibility of predicting gene expression in *Plasmodium*, we next turned to the more interesting question: which features contribute most strongly to the performance of each classifier? By including three different types of models, we reasoned that if multiple classification models select a similar set of informative features within a single stage, then this would suggest that those features are robust to the choice of model. Accordingly, for each model we calculated a feature importance score (see Methods) on a 0 to 1 scale, where 0 means uninformative and 1 means strongly informative. We also determined the direction of effect, indicating whether a high feature value is predictive of high or low expression. Additionally, the DeepPINK model identifies which features are informative for classification using a method that allows for explicit control of false discovery rate (FDR $< 0.05$, see Methods). Note that due to the exclusion of motif features from our analysis (Fig 2C), we do not obtain feature scores for any AP2 motifs. Given the fundamentally different methods used to assign feature importances in each of the three models, we would not expect a precise quantitative agreement in scores. For instance, elastic-net regularized regression models favor zero-valued coefficients, whereas no such sparsity-inducing behavior occurs in XGBoost or DeepPINK models. However, we would expect that model agreement on feature importances would result in similar feature rankings. Indeed, when we calculate the Spearman correlation between all model pairings in a given stage we see a strong correlation between different models' feature orderings (Table 4). We note that the agreement is imperfect, particularly between XGBoost and DeepPINK models in the schizont stage (correlation = 0.533), with obvious differences between the two models in their use of H3K36me2 and H3K36me3, among others (Fig 4A). In general, occasional instances of disagreement between models feature attributions within a given stage (Fig 4A and Table 4) are difficult to interpret with confidence given the lack of a ground truth to which we can compare the models' feature attributions.

**Table 4. Intra-stage consistency of model feature attributions.** Spearman correlations between all pairs of models were calculated for features within individual stages, as well as averaged across all stages.

| Model comparison | Ring features | Trophozoite features | Schizont features |
|---|---|---|---|
| Logistic vs. XGBoost | 0.945 | 0.934 | 0.718 |
| Logistic vs. DeepPINK | 0.876 | 0.857 | 0.772 |
| XGBoost vs. DeepPINK | 0.821 | 0.775 | 0.533 |

https://doi.org/10.1371/journal.pcbi.1007329.t004

Given the observation that our models attribute similar importances to features within a single stage (Table 4 and Fig 4A), we inspected the features that the models selected as informative. All three models identify high H3K4me3 signal within the gene body as indicative of high expression in the ring stage, select high H3K4me3 signal as indicative of high expression in the trophozoite stage, and identify high H3K9Ac and H4K20me3 signal in the gene body as indicative of high expression in the schizont stage. Similarly, the three models tend to attribute consistent importance to physical chromatin features: all three models highlight the importance of gene body nucleosome occupancy in the trophozoite stage and telomere distance in the ring stage (using inferred distance based on a 3D computational model, see Methods). This consistency across models and methods suggests that our approach to identifying informative features is generally robust to the differences in modeling approaches.

In contrast to the high concordance among the three models, we observed low concordance among the importance of individual features across different stages of the erythrocytic cycle. For instance, H4K20me3 was highly informative for predicting a "high expression" label in the schizont stage, but almost completely uninformative in the ring stage (data was unavailable for this mark in the trophozoite stage). To investigate the extent of this disagreement, we calculated Spearman correlations for all stage pairs for a given model type (Table 5). These correlations (mean = 0.623) are notably lower than the correlations observed between different models trained within the same stage (Table 4, mean = 0.803). The comparatively higher consistency of feature importance within a stage versus between stages (difference = .18. p = .0176, two-sided t-test) argues that inter-stage differences are not an artifact of the model training process and suggests that distinct regulatory mechanisms may control transcription in the three different stages. However, further work and replicatation is required to rule out confounding issues such as batch effects between datasets for different stages. For instance, the H3K4me3 features from one source [35] was marked as informative for classification in the schizont stage by all three models, while H3K4me3 signal from a different source [33] was found to be relatively uninformative, by comparison (Fig 4A). Such discrepancies likely stem from differences in either the data generation processes or the synchronization of parasitic stages across distinct sources.

The XGBoost model afforded an additional look at each individual features' effects on the classification of single genes. In addition to assigning feature significance and direction-of-effect at the level of the model as a whole (as in Fig 4A), the SHAP score for the XGBoost model can be calculated separately for each feature at each gene. Briefly, to generate classifications the XGBoost model generates a score for each gene. Suppose that correctly classified "low-expression" genes receive scores in the range 2–8 (on an arbitrary scale), and a particular low-expression gene has been given a score of 6, indicating that the model (correctly) predicts that the gene is at a low expression level. SHAP scores assign scores to each of the 23 features used in the prediction, such that the sum of the 23 scores adds up to 6, the final classification value. In this way, large positive SHAP scores indicate features that were important for assigning this particular gene a "high-expression" label (see Section and [51] for details). The resulting distribution of per-gene SHAP scores for each feature (Fig 4B) suggests that some features

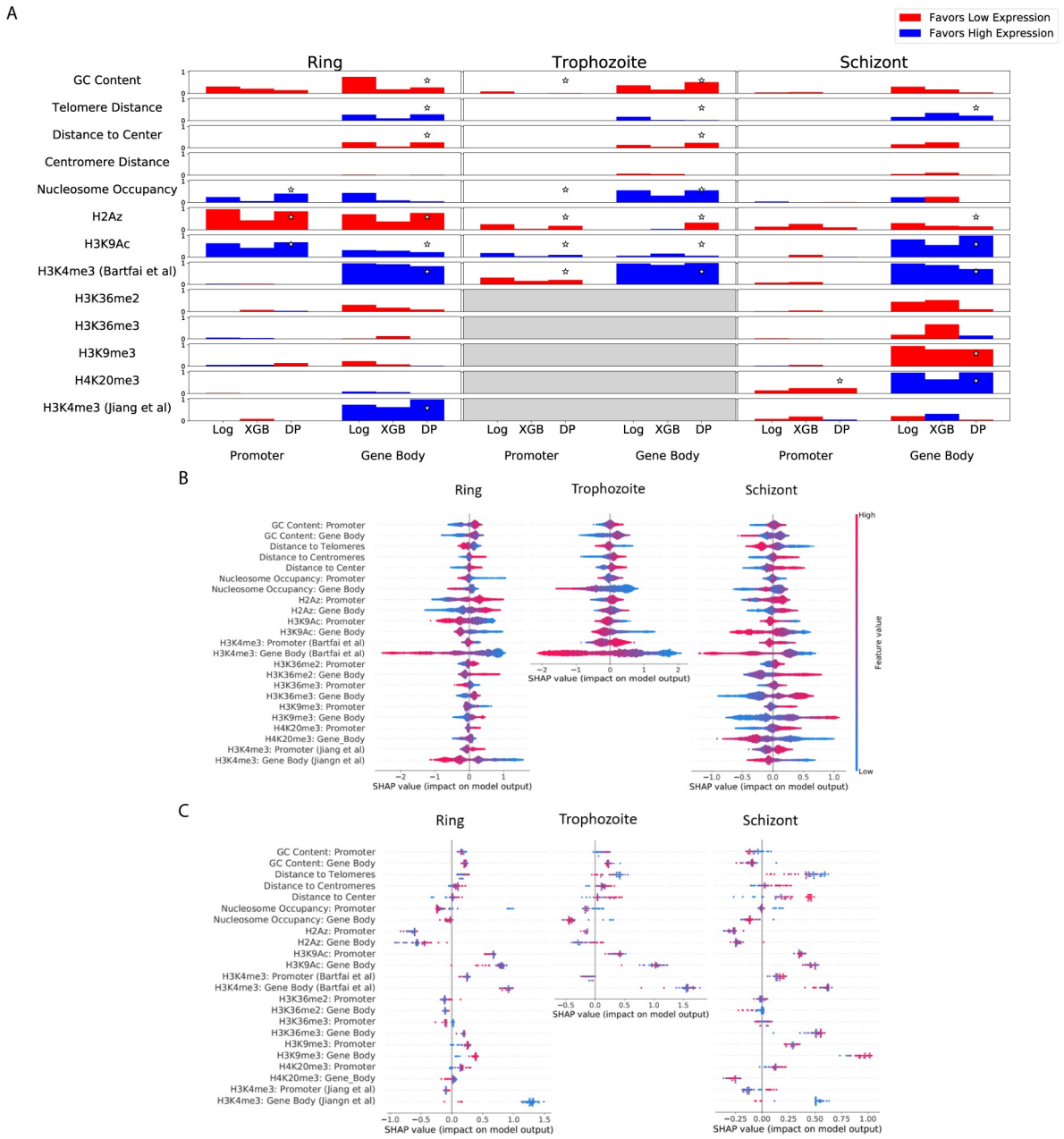


**Fig 4. Feature importance measures.** (A) Feature importance scores assigned for different models (Log = Logistic Regression, XGB = XGBoost, DP = DeepPINK Multilayer Perceptron). See "Methods" for details on how feature importance scores were calculated. Scores were normalized to lie within a 0-to-1 range by subtracting the minimum absolute value of all scores in a model, then dividing those numbers by the score with the maximum absolute value. Bar height represents the magnitude of feature significance, while the color of bars indicates the direction of effect (Red: Higher feature value predicts high expression. Blue: Higher feature value predicts low expression.). Features using averages over "promoter" and "gene body" windows (such as ChIP-seq tracks) are split by these sub-features, while features that are not divided (such as distance to centromere centroid) are not. Stars indicate features that were selected as significant using the DeepPINK model, controlling false discovery rate $<0.05$. (B) SHAP values for features used in the XGboost classifier for all genes. SHAP values for a given gene represent how significant a specific feature was for classification of a gene as low-expression or high-expression, as well as the direction in which the feature pushed the classification. A positive SHAP score for a feature for a specific gene means that the value of that feature was changed that gene's classification toward "low expression," while a negative SHAP value means that feature pushed the gene toward a label of "high expression." (C) SHAP values for features used in the XGboost classifier for virulence genes.

https://doi.org/10.1371/journal.pcbi.1007329.g004

**Table 5. Inter-stage consistency of model feature attributions.** Spearman correlations between all pair-wise comparisons of stages were calculated for features within individual model types.

| Model | Ring vs. Trophozoite | Ring vs. Schizont | Trophozoite vs. Schizont |
|---|---|---|---|
| Logistic | .830 | .632 | .722 |
| XGB | .711 | .641 | .365 |
| DeepPINK | .758 | .385 | .565 |

https://doi.org/10.1371/journal.pcbi.1007329.t005

exhibit non-linear relationships between SHAP scores and expression prediction, visually observable as asymmetry in the density plots shown in Fig 4B. For instance, the effect of H3K36me2 gene body signal in the schizont stage model is not a simple relationship where increases in the feature value lead to consistent changes in model classification Fig 5A. In a histogram showing H3K36me2 distributions for high- and low-expression genes (Fig 4C), we see that over the range of -4 to -3, almost all genes are "low-expression." This corresponds to the "flat" region in the -4 to -3 range in the SHAP scores of Fig 4A, because values within this range are all treated as essentially the same by the XGBoost model. In contrast, between 0 and 1 we observe a shift in the relative abundances of high- and low-expression genes: most genes with 0 signal are high-expression, whereas most genes with 1 signal are low expression.

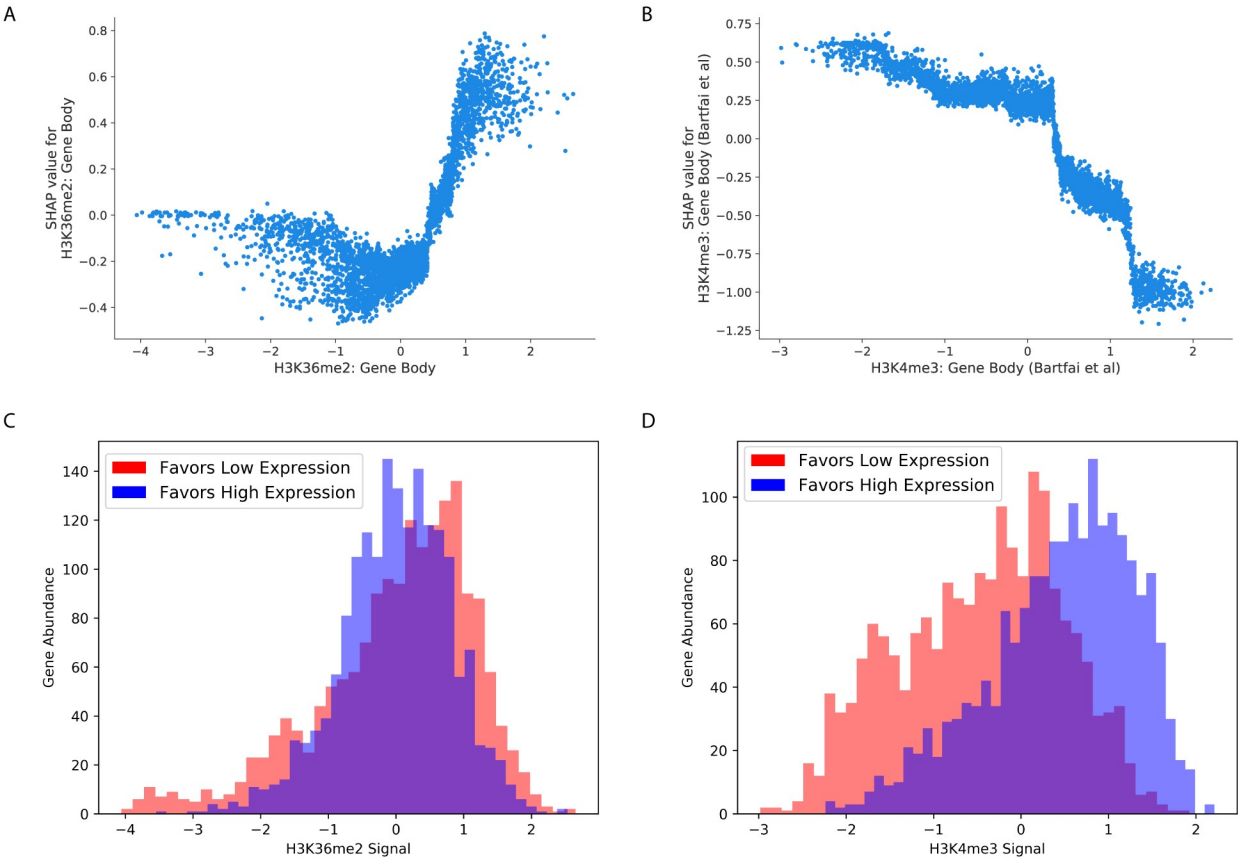


**Fig 5. Nonlinear relationships between feature values and SHAP values.** (A) Scatter plot showing the values of H3K36me2 signal in the gene body of schizont-stage genes (x-axis) against the SHAP values assigned to those genes (y-axis). (B) A scatter plot plot showing H3K4me3 in the gene body of schizont genes ([35], x-axis) against SHAP values for those genes (y-axis). (C) Histogram showing the H3K36me2 feature value distributions for low-expression (red) and high-expression (blue) genes. (D) Histogram showing H3K4me3 values for low-expression (red) and high-expression (blue) genes.

https://doi.org/10.1371/journal.pcbi.1007329.g005

Consequently, we see that SHAP scores have a steep slope over the 0 to 1 range, meaning that small changes in H3K36me2 have large effects on model predictions for genes within this range. Similarly, a unit change in gene body H3K4me3 intensity does not lead to a specific change in XGBoost predictions across all ranges of H3K4me3 signal (Fig 5B). We see a marked change in the relationship (slope) between SHAP scores and H3K4me3 signal around a score of "0.5" (Fig 5B), which is at the point at which genes transition from being mostly low-expression to mostly high-expression, as seen in the distributions of Fig 5D.

Intuitively, this observation indicates that the XGBoost model can discriminate between feature input ranges where small changes are important versus regions where small changes are insignificant. This can help capture the behavior of underlying non-linear mechanisms: for instance, high levels of H3K4me3 may indirectly help recruit pre-initiation complex components, but after a certain level H3K4me3-mediated recruitment is no longer the rate-limiting step for transcription, so further H3K4me3 deposition will not further increase polymerase activity. Non-linear models such as XGBoost and DeepPINK are able to capture such feature-response nonlinearities, allowing for improved predictions when modeling a process with significantly non-linear mechanisms. In contrast, a linear model like logistic regression treats -4 and -3 as being exactly as different as 0 and 1, regardless of the underlying feature distribution.

We also repeated the per-gene SHAP analysis for the *Plasmodium* virulence genes, which encode a protein family that functions to anchor infected erythrocytes to the endothelium of blood vessels and are an important target for immune recognition [53]. The virulence genes are tightly regulated, with each parasite expressing exactly one of the 60 genes at a given time. In agreement with a known role for H3K9me3 in repression of virulence genes [7], we find that virulence genes have large SHAP scores assigned to H3K9me3 signal. This observation demonstrates that the classification model is not only able to find genome-wide rules for classification, but also selects important features with respect to a specific subset of genes, capturing factors that are known to be important for transcriptional control of that gene family.

## Discussion

We developed predictive models for *Plasmodium* gene expression that yield AUC values in the range of 0.79–0.88 in cross-validated testing. These values are somewhat lower than AUC values reported from studies carried out in other eukaryotes like mouse (0.94 [15]) or human (0.95 [16]). Many factors may contribute to this difference. For example, *Plasmodium* has a smaller number of datasets available for use as features in our models: at most six unique histone covalent modifications were used in our models, whereas 11 unique histone modification features were used in both [15] and [16]. Consistent with this, using a small feature set (five histone modifications) to classify expression in human cells resulted in a model with an average AUC of $\sim$0.8, a value in line with the performance we observed. Furthermore, compared to human and mouse, *Plasmodium* has far fewer genes, which yields fewer examples for training our models. Additionally, the high AT-content of the *Plasmodium* genome presents a consistent challenge to generation of high-confidence genomic datasets [9], so noise in feature datasets may have led to reduced accuracy. An alternate explanation comes from the apparent abundance of genes related to post-transcriptional regulation, rather than gene-specific transcriptional control [2]. This discrepancy has led to speculation that the most significant level of gene expression control occurs at regulation of translation, relaxing requirements for strict transcriptional regulatory programs [2]. It is possible that a relatively low reliance on strict transcriptional control allows the parasite to tolerate high noise in transcriptional regulation, in turn leading to a system that is difficult to model accurately.

One surprising outcome was the apparently low utility of features derived from AP2 family TF binding motifs. *Plasmodium* AP2 genes are conserved proteins containing putative DNA-binding domains, homologous to the plant Apetela2/Ethylene Response Factor (AP2/ERF) DNA-binding proteins, the second largest class of transcription factors in *Arapidopsis thaliana* ([54]). Gene expression profiling of a number of *Plasmodium* species as well as targeted knock-out studies have demonstrated that some of these proteins are transcriptionally regulated and play key roles during developmental stages, including sexual differentiation ([55]). Finally, a pronounced paucity of alternative transcription factors with DNA sequence specificity [43] and the presence of high-affinity AP2 motifs upstream of genes whose expression varied across the erythrocytic cycle led to the notion that these Ap2 factors could be the missing reservoir of sequence specific TFs in *Plasmodium*. The strikingly low value of AP2 motifs that we obtained raises three possibilities. First, this result may be indicative of the relatively low importance of local TF activity in regulating gene expression during the erythrocytic cycle. Second, we cannot completely rule out the possibility that the low value of AP2 motifs arose simply because the motifs used here are of low quality or because the way we employed the motifs (by scanning and aggregating p-values) is suboptimal. Alternately, it is possible that AP2 DNA binding requires both a TF-specific motif and a permissive epigenetic state at a given locus. DNA accessibility and epigenetic state is known to play a role in restricting TF binding in eukaryotes generally [56], with the consequence that TF motifs are an imperfect predictor of DNA binding in the absence of additional epigenomic data [30, 57]. In *Plasmodium* specifically, detailed study of one TF found that the presence of a consensus motif was neither strictly necessary nor sufficient for TF binding [58]. If local chromatin state affects TF binding even in the presence of a TF-specific DNA motif, the predictive value of AP2 motifs could be masked by subtle interactions with local DNA accessibility and chromatin state. Consistent with this possibility, previous models of mammalian gene expression based on sequence motifs captured a small amount of gene expression variation [18, 59], while models using TF binding assayed by ChIP-Seq were able to predict expression with far greater accuracy [60]. This is presumably because ChIP-Seq data implicitly captures both motif presence/absence as well as epigenetic factors affecting TF binding. Clearly, an extensive collection of TF ChIP-seq data would be hugely valuable in exploring the extent to which TFs play an active role in gene regulation in *Plasmodium* and would clarify if AP2 factors truly play a limited role in erythrocytic transcriptional regulation. Initial Chip-seq results against AP2-G2 and AP2-I, transcription factors thought to be involved in sexual development and cell invasion respectively, suggest that AP2 may interact with some promotors to either act as a repressor for AP2-G2 ([61]) or activator in association with several chromatin-associated proteins, including the *Plasmodium* bromodomain protein PfBDP1 for Ap2I ([62]).

Inspection of our trained models revealed the use of multiple types of features, from local histone modifications to high-order spatial positioning. Covalent histone modifications were consistently found to be informative features, including the designation of gene-body H3K9Ac and H3K4me3 [35] as statistically significant by DeepPINK FDR control in all three stages (Fig 4A). Furthermore, nucleosome occupancy and GC content were repeatedly identified as informative features (Fig 4A, Ring and Trophozoite feature use). Together, these observations indicate that nucleosome occupancy, histone modification status, and GC content all contain valuable information regarding the activity status of a locus. In addition, the gene distances to telomere cluster and nuclear center (based on 3D models from our groups' previously generated data, see Methods) were also consistently informative for classification of *Plasmodium* gene expression, albeit to a lesser extent than local features such as histone modifications (Fig 4A). This is consistent with previous observations that *Plasmodium* expression correlates with

gene spatial positioning [34], and suggests that *Plasmodium* may encode regulatory information in the 3D position of a gene, in addition to its local epigenetic state. Our findings complement previous identification of co-regulatory relationships between functionally related genes in *Plasmodium* [63], with our analysis identifying a repertoire of epigenetic features that underpin such observed patterns.

Interestingly, in the DeepPINK model H2Az coverage in the gene body of trophozoite genes was marked as significant (FDR $< 0.05$) and associated with low expression. In contrast, scores assigned to this feature were close to zero for both the Logistic and XGBoost models. H2Az signal was previously reported to be almost completely absent from gene coding sequence [35], which makes the apparent significance of gene-body H2Az signal quite surprising. Follow-up validation would be required to see if a minimal level of H2Az truly encodes information within coding sequence, or if the identification of the feature as significant is an artifact of the DeepPINK procedure. However, previous studies in metazoan genomes have also identified H2A.Z in gene bodies. While some research groups link low levels of H2A.Z with inhibition of transcription in reconstituted nucleosomes [64, 65], others suggest that H2A.Z nucleosomes may facilitate transcriptional elongation [66]. Our results support a model in which a low level of H2A.Z nucleosomes acts as a simple barrier to transcriptional elongation. However, given the general agreement between models for almost all other features (Fig 4A) the assignment of importance to H2Az signal by DeepPINK alone suggests that the relationship should be considered very tentative.

An inherent limitation of our analysis is that, given these data, we cannot easily separate correlations from causative relationships. This is particularly important when modeling transcription using epigenetic data, given previous evidence that some histone marks (H3K36 and H3K79 methylation) are deposited directly as an effect of Pol II elongation, rather than preceding transcriptional activation [67]. In the absence of detailed perturbational experiments, the predictive relationships that we observe between features and expression cannot be clearly defined as directly regulatory or not.

Despite this limitation, our identification of predictive features is helpful on two fronts. First, epigenomic changes resulting from transcriptional activity can themselves serve in regulatory roles. In some species, H3K36 methylation, for instance, is deposited concurrently with transcription but serves a regulatory role thereafter, suppressing aberrant initiation of transcription within gene bodies [67]. This means that our models may identify factors important not only for regulation preceding initial activation of a locus, but also for feedback regulatory mechanisms. Second, the observed differences in selected features in distinct stages gives a clear prioritization for points in the *Plasmodium* life cycle where experimental dissection of epigentic function would be most informative. For instance, H4K20me3 is not predictive of expression the ring stage, but is consistently associated with transcriptional repression in the schizont stage (Fig 4A). Whether H4K20me3 is a cause or effect of transcription, the molecular events linking this mark to transcription likely only take place—and are experimentally targetable—in the schizont stage. Our analysis specifically suggests that H3K9me3, H4K20me3, and K3K9Ac play schizont-specific regulatory roles in the erythrocytic cycle of *Plasmodium* (Fig 4A). This observation suggests that disruption of enzymes controlling the levels of these marks would result in schizont-specific disregulation, either through genetic ablation or chemical inhibition. Therapeutic targetting of specific epigenetic pathways is already an active area of study in oncology [68] and virology [69], and future efforts applying epigenetic disruption to antimalarial regimens will benefit from our determination that the schizont stage appears to rely upon a larger variety of covalent histone modifications than other erythrocytic stages.

Analyzing XGBoost models suggested that the best solution to the classification task did not take the simple form in which a unit increase in a given feature leads to a specific, constant

change in classification probability (Figs 4B and 5). Consistent with this, our two approaches that allow for feature interactions and non-linear feature/classification relationships, XGBoost and DeepPINK, slightly but consistently out-performed logistic regression (Fig 2D). However, the improvements in test AUC for the XGBoost/DeepPINK models are statistically significant in only a subset of these comparisons, and in all cases are quite modest in absolute terms. This is consistent with work in other eukaryotic genomes, where incorporating feature interactions provided minimal improvement in gene expression prediction accuracy, compared to simple linear models [15, 16]. It is possible that complex models would demonstrate a more significant advantage in a regression task—such as predicting absolute mRNA abundance—rather than the binary classification task that we considered. In our case, however, it appears that models using simple additive effects, such as logistic regression, captured most of the information found within the input features.

Our work only studied factors associated with relative control of expression within a particular erythrocytic stage. This approach has the limitation of ignoring gene expression dynamics related to changes in absolute expression. In our per-stage labeling approach, 2937 genes received the same label ("high", "low", or "intermediate") in all three stages, 2162 were either "high" and "intermediate" or "low" and "intermediate", and the remaining 182 were labeled as "high" and "low" expression at least once. Fig 1A shows that genome-wide expression values vary widely between stages, consistent with known inter-stage variation in transcriptional activity. From this, we know that many genes are changing in absolute expression levels between stages, but ending up with a "constant" expression label when classified by relative expression. Conversely, a subset of housekeeping or otherwise constantly expressed genes may not actually vary in absolute expression themselves but end up with varying high/intermediate/low labels due to global shifts of transcription. An alternative modeling approach could build upon our analysis of intra-stage expression regulation to incorporate inter-stage expression changes and absolute expression regulation, with the aim of building a complementary picture of *Plasmodium* transcriptional control.

During model development and feature refinement, we came to the surprising discovery that placing the promoter/gene body division using start codon position was more effective than using transcription start sites (Fig 3B). This observation is consistent with a previous analysis in which five out of six covalent histone modifications associated with high transcription in *Plasmodium* displayed peak enrichment at the start codons of *Plasmodium* genes, while only one displayed the highest enrichment at transcription start sites [70]. Additionally, this is consistent with the observation that *Plasmodium* lacks a strongly positioned +1 nucleosome at the TSS, but that clearly positioned nucleosomes are observed at the start and end of coding sequences [8, 9]. In the future, it would be interesting to see if epigenetic information related to transcriptional control is truly encoded primarily with respect to start codons, or if technical artifacts due to the extreme AT bias in non-coding DNA upstream of start codons leads to the apparently limited information value of TSS-centered signals.

## Supporting information

**S1 Fig. Varying promoter sizes.** Comparison of test AUCs for motif-only models using 500 bp, 1kb, or 2kb promoter windows.
(PDF)

**S1 Table. GRO-seq data.** GRO-seq values, normalized for GC content, gene length and the parasitemia factor of a stage [37].
(XLSX)

## Author Contributions

**Conceptualization:** David F. Read, Kate Cook, Karine G. Le Roch, William Stafford Noble.

**Data curation:** Kate Cook, Karine G. Le Roch.

**Investigation:** David F. Read.

**Methodology:** Kate Cook, Yang Y. Lu.

**Supervision:** Karine G. Le Roch, William Stafford Noble.

**Validation:** David F. Read.

**Visualization:** David F. Read.

**Writing – original draft:** David F. Read, Kate Cook, Karine G. Le Roch, William Stafford Noble.

**Writing – review & editing:** David F. Read, Kate Cook, Yang Y. Lu, Karine G. Le Roch, William Stafford Noble.

## References


1. Organization TWH. World malaria report 2017.
2. Coulson RM, Hall N, Ouzounis CA. Comparative genomics of transcriptional control in the human malaria parasite *Plasmodium falciparum*. Genome Research. 2004; 14(8):1548–1554. https://doi.org/10.1101/gr.2218604 PMID: 15256513
3. de Boer CG, Hughes TR. YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. Nucleic Acids Research. 2012. https://doi.org/10.1093/nar/gkr993
4. De Silva EK, Gehrke AR, Olszewski K, Leon I, Chahal JS, Bulyk ML, et al. Specific DNA binding by apicomplexan AP2 transcription factors. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105(24):8393–8398. https://doi.org/10.1073/pnas.0801993105 PMID: 18541913
5. Kafsack BF, Rovira-Graells N, Clark TG, Bancells C, Crowley VM, Campino SG, et al. A transcriptional switch underlies commitment to sexual development in malaria parasites. Nature. 2014; 507(7491):248–252. https://doi.org/10.1038/nature12920 PMID: 24572369
6. Sinha A, Hughes KR, Modrzynska KK, Otto TD, Pfander C, Dickens NJ, et al. A cascade of DNA-binding proteins for sexual commitment and development in *Plasmodium*. Nature. 2014; 507(7491):253–257. https://doi.org/10.1038/nature12970 PMID: 24572359
7. Chookajorn T, Dzikowski R, Frank M, Li F, Jiwani AZ, Hartl DL, et al. Epigenetic memory at malaria virulence genes. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104(3):899–902. https://doi.org/10.1073/pnas.0609084103 PMID: 17209011
8. Bunnik EM, Polishko A, Prudhomme J, Ponts N, Gill SS, Lonardi S, et al. DNA-encoded nucleosome occupancy is associated with transcription levels in the human malaria parasite *Plasmodium falciparum*. BMC Genomics. 2014; 15:347. https://doi.org/10.1186/1471-2164-15-347 PMID: 24885191
9. Ay F, Bunnik EM, Varoquaux N, Vert JP, Noble WS, Le Roch KG. Multiple dimensions of epigenetic gene regulation in the malaria parasite *Plasmodium falciparum*. Bioessays. 2015; 37(2):182–194. https://doi.org/10.1002/bies.201400145 PMID: 25394267
10. Lopez-Rubio JJ, Mancio-Silva L, Scherf A. Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. Cell. 2009; 5:179–190.
11. Petter M, Selvarajah SA, Lee CC, Chin WH, Gupta AP, Bozdech Z, et al. H2A.Z and H2B.Z double-variant nucleosomes define intergenic regions and dynamically occupy var gene promoters in the malaria parasite *Plasmodium falciparum*. Mol Microbiol. 2013; 87(6):1167–1182. https://doi.org/10.1111/mmi.12154 PMID: 23373537
12. Ponts N, Harris EY, Prudhomme J, Wick I, Eckhardt-Ludka C, Hicks GR, et al. Nucleosome landscape and control of transcription in the human malaria parasite. Genome Research. 2010; 20(2):228–238. https://doi.org/10.1101/gr.101063.109 PMID: 20054063
13. Toenhake CG, Fraschka SA, Vijayabaskar MS, Westhead DR, van Heeringen SJ, Bártfai R. Chromatin accessibility-based characterization of the gene regulatory network underlying *Plasmodium falciparum*

blood-stage development. Cell Host Microbe. 2018; 23(4). https://doi.org/10.1016/j.chom.2018.03.007 PMID: 29649445

14. Bunnik EM, Cook KB, Varoquaux N, Batugedara G, Prudhomme J, Cort A, et al. Changes in genome organization of parasite-specific gene families during the *Plasmodium* transmission stages. Nature Communications. 2018; 15(9):1910. https://doi.org/10.1038/s41467-018-04295-5
15. Cheng C, Yan KK, Yip KY, Rozowsky J, Alexander R, Shou C, et al. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. Genome Biology. 2011; 12(2):R15. https://doi.org/10.1186/gb-2011-12-2-r15 PMID: 21324173
16. Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, et al. Modeling gene expression using chromatin features in various cellular contexts. Genome Biology. 2012; 13. https://doi.org/10.1186/gb-2012-13-9-r53
17. Singh R, Lanchantin J, Robins G, Qi Y. Deepchrome: deep-learning for predicting gene expression from histone modifications. Bioinformatics. 2016; 32(17):i639–i649. https://doi.org/10.1093/bioinformatics/btw427 PMID: 27587684
18. Bussemaker HJ, Li H, Siggia ED. Regulatory element detection using correlation with expression. Nature Genetics. 2001; 27:167–171. https://doi.org/10.1038/84792 PMID: 11175784
19. Pe'er D, Regev A, Tanay A. Minreg: inferring an active regulator set. Bioinformatics. 2002; 18:S258–67. https://doi.org/10.1093/bioinformatics/18.suppl_1.s258 PMID: 12169555
20. Segal E, Yelensky R, Koller D. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. Bioinformatics. 2003; 19(Suppl. 1):i273–i282. https://doi.org/10.1093/bioinformatics/btg1038 PMID: 12855470
21. Beer MA, Tavazoie S. Predicting gene expression from sequence. Cell. 2004; 117:185–198. https://doi.org/10.1016/s0092-8674(04)00304-6 PMID: 15084257
22. Middendorf M, Kundaje A, Wiggins C, Freund Y, Leslie C. Predicting genetic regulatory response using classification. Bioinformatics. 2004; 20:i232–40. https://doi.org/10.1093/bioinformatics/bth923 PMID: 15262804
23. Kundaje A, Middendorf M, Shah M, Wiggins CH, Freund Y, Leslie C. A classification-based framework for predicting and analyzing gene regulatory response. BMC Bioinformatics. 2006; 7, Suppl 1. https://doi.org/10.1186/1471-2105-7-S1-S5 PMID: 16723008
24. Ouyang Z, Zhou Q, Wong HW. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106:21521–21526. https://doi.org/10.1073/pnas.0904863106 PMID: 19995984
25. Karlic R, Chung HR, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are predictive for gene expression. Proceedings of the National Academy of Sciences of the United States of America. 2010; 107(7):2926–2931. https://doi.org/10.1073/pnas.0909344107 PMID: 20133639
26. McLeay RC, Lesluyes T, Partida GC, Bailey TL. Genome-wide in silico prediction of gene expression. Bioinformatics. 2012; 28:2789–2796. https://doi.org/10.1093/bioinformatics/bts529 PMID: 22954627
27. Zhou X, Cain CE, Myrthil M, Lewellen N, Michelini K, Davenport ER, et al. Epigenetic modifications are associated with inter-species gene expression variation in primates. Genome Biology. 2014; 15. https://doi.org/10.1186/s13059-014-0547-3
28. González AJ, Setty M, Leslie CS. Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation. Nature Genetics. 2015; 47:1249–59. https://doi.org/10.1038/ng.3402 PMID: 26390058
29. Chen Y, Li Y, Narayan R, Subramanian A, Xie X. Gene expression inference with deep learning. Bioinformatics. 2016; 32(12):1832–1839. https://doi.org/10.1093/bioinformatics/btw074 PMID: 26873929
30. Duren Z, Chen X, Jiang R, Wang Y, Wong WH. Modeling gene regulation from paired expression and chromatin accessibility data. Proceedings of the National Academy of Sciences of the United States of America. 2017; 114. https://doi.org/10.1073/pnas.1704553114 PMID: 28576882
31. Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res. 2018. https://doi.org/10.1101/gr.227819.117 PMID: 29588361
32. Osmanbeyoglu HU, Shimizu F, Rynne-Vidal A, Jelinic P, Mok SC, Chiosis G, et al. Chromatin-informed inference of transcriptional programs in gynecologic and basal breast cancers. bioRxiv. 2018.
33. Jiang L, Mu J, Zhang Q, Ni T, Srinivasan P, Rayavara K, et al. PfSETvs methylation of histone H3K36 represses virulence genes in *Plasmodium falciparum*. Nature. 2013; 499(7457):223–227. https://doi.org/10.1038/nature12361 PMID: 23823717
34. Ay F, Bunnik EM, Varoquaux N, Bol SM, Prudhomme J, Vert JP, et al. Three-dimensional modeling of the *P. falciparum* genome during the erythrocytic cycle reveals a strong connection between genome

architecture and gene expression. Genome Research. 2014; 24:974–988. https://doi.org/10.1101/gr.169417.113 PMID: 24671853

35. Bartfai R, Hoeijmakers WA, Salcedo-Amaya AM, Janssen-Megens AHSE, Kaan A, Treeck M, et al. H2A.Z demarcates intergenic regions of the *Plasmodium falciparum* epigenome that are dynamically marked by H3K9ac and H3K4me3. PLOS Pathogens. 2010; 6(12):e1001223. https://doi.org/10.1371/journal.ppat.1001223 PMID: 21187892

36. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014; 158(6):1431–1443. https://doi.org/10.1016/j.cell.2014.08.009 PMID: 25215497

37. Lu XM, Batugedara G, Lee M, Prudhomme J, Bunnik EM, Le Roch KG. Nascent RNA sequencing reveals mechanisms of gene regulation in the human malaria parasite *Plasmodium falciparum*. Nucleic Acids Research. 2017; 45(13):7825–7840. https://doi.org/10.1093/nar/gkx464 PMID: 28531310

38. Adjalley SH, Chabbert CD, Klaus B, Pelechano V, Steinmetz LM. Landscape and dynamics of transcription initiation in the malaria parasite *Plasmodium falciparum*. Cell Reports. 2016; 14(10). https://doi.org/10.1016/j.celrep.2016.02.025 PMID: 26947071

39. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010; 26(5):589–595. https://doi.org/10.1093/bioinformatics/btp698 PMID: 20080505

40. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009; 25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

41. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26(6):841–842. https://doi.org/10.1093/bioinformatics/btq033 PMID: 20110278

42. Varoquaux N, Ay F, Noble WS, Vert JP. A statistical approach for inferring the 3D structure of the genome. Bioinformatics. 2014; 30(12):i26–i33. https://doi.org/10.1093/bioinformatics/btu268 PMID: 24931992

43. Campbell TL, de Silva EK, Olszewski KL, Elemento O, Llinas M. Identification and Genome-Wide Prediction of DNA Binding Specificities for the ApiAP2 Family of Regulators from the Malaria Parasite. PLOS Pathogens. 2010; 6(10):e1001165. https://doi.org/10.1371/journal.ppat.1001165 PMID: 21060817

44. Balaji S, Babu MM, Iyer LM, Aravind L. Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. Nucleic Acids Research. 2005; 33(13):3994–4006. https://doi.org/10.1093/nar/gki709 PMID: 16040597

45. Grant CE, Bailey TL, Noble WS. FIMO: Scanning for occurrences of a given motif. Bioinformatics. 2011; 27(7):1017–1018. https://doi.org/10.1093/bioinformatics/btr064 PMID: 21330290

46. Bailey T, Boden M, Buske F, Frith M, Grant CE, Clementi L, et al. MEME Suite: Tools for motif discovery and searching. Nucleic Acids Research. 2009; 37(Web server issue):W202–208. https://doi.org/10.1093/nar/gkp335 PMID: 19458158

47. Lu YY, Fan Y, Lv J, Noble WS. DeepPINK: reproducible feature selection in deep neural networks. In: Advances in Neural Information Processing Systems; 2018.

48. DeLong ER, Delong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988. https://doi.org/10.2307/2531595 PMID: 3203132

49. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011. https://doi.org/10.1186/1471-2105-12-77 PMID: 21414208

50. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD'16. New York, NY, USA: ACM; 2016. p. 785–794. Available from: http://doi.acm.org/10.1145/2939672.2939785.

51. Lundberg SM, Lee S. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. 2017.

52. Barber RF, Candes EJ. Controlling the false discovery rate via knockoffs. The Annals of Statistics. 2015; 43(5):2055–2085. https://doi.org/10.1214/15-AOS1337

53. Flick K, Chen Q. *var* genes, PfEMP1 and the human host. Mol Biochem Parasitol. 2004; 134(1). https://doi.org/10.1016/j.molbiopara.2003.09.010 PMID: 14747137

54. Riechmann JL, Meyerowitz EM. The AP2/EREBP family of plant transcription factors. Biol Chem. 1998; 379. PMID: 9687012

55. Painter HJ, Campbell TL, Llinás M. The apicomplexan AP2 family: integral factors regulating Plasmodium development. Mol Biochem Parasitol. 2010; 176. https://doi.org/10.1016/j.molbiopara.2010.11.014 PMID: 21126543

56. MacQuarrie KL, Fong AP, Morse RH, Tapscott SJ. Genome-wide transcription factor binding: beyond direct target regulation. Trends Genet. 2011; 27. https://doi.org/10.1016/j.tig.2011.01.001 PMID: 21295369

57. Blatti C, Kazemian M, Wolfe S, Brodsky M, Sinha S. Integrating motif, DNA accessibility and gene expression data to build regulatory maps in an organism. Nucleic Acids Res. 2015; 43. https://doi.org/10.1093/nar/gkv195 PMID: 25791631

58. Gissot M, Briquet S, Refour P, Boschet C, Vaquero C. PfMyb1, a Plasmodium falciparum transcription facator, is required for intra-erythrocytic growth and controls key genes for cell cycle regulation. J Mol Biol. 2005; 11.

59. Conlon EM, Liu XS, Lieb JD, Liu JS. Integrating regulatory motif discovery and genome-wide expression analysis. Proceedings of the National Academy of Sciences of the United States of America. 2003; 100(6):3339–3344. https://doi.org/10.1073/pnas.0630591100 PMID: 12626739

60. Cheng C, Gerstein M. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. Nucleic Acids Research. 2012; 40(2):553–568. https://doi.org/10.1093/nar/gkr752 PMID: 21926158

61. Yuda M, Iwanaga S, Kaneko I, Kato T. Global transcriptional repression: An initial and essential step for Plasmodium sexual development. PNAS. 2015; 112. https://doi.org/10.1073/pnas.1504389112 PMID: 26417110

62. Santos JM, Josling G, Ross P, Joshi P, Orchard L, Campbell T, et al. Red blood cell invasion by the malaria parasite is coordinated by the PfAP2-I transcription factor. Cell Host Microbe. 2017; 21. https://doi.org/10.1016/j.chom.2017.05.006

63. Prat Y, Fromer M, Linial N, Linial M. Recovering key biological constiituents through sparse representation of gene expression. Bioinformatics. 2011; 27(5). https://doi.org/10.1093/bioinformatics/btr002 PMID: 21258061

64. Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, Zanton SJ, et al. Nucleosome organization in the *Drosophila* genome. Nature. 2008; 453(7193):358–362. https://doi.org/10.1038/nature06929 PMID: 18408708

65. Thakar A, Gupta P, McAllister WT, Zlatanova J. Histone variant H2A.Z inhibits transcription in reconstituted nucleosomes. Biochemistry. 2010; 49(19):4018–4026. https://doi.org/10.1021/bi1001618 PMID: 20387858

66. Weber CM, Henikoff JG, Henikoff S. H2A.Z nucleosomes enriched over active genes are homotypic. Nature Structural and Molecular Biology. 2010; 17:1500–1507. https://doi.org/10.1038/nsmb.1926 PMID: 21057526

67. Gates LA, Foulds CE, O'Malley BW. Histone marks in the 'Driver's Seat': Functional roles in steering the transcriptional cycle. Trends in Biochemical Sciences. 2017; 42. https://doi.org/10.1016/j.tibs.2017.10.004 PMID: 29122461

68. Fardi M, Solali S, Hagh MF. Epigenetic mechanisms as a new approach in cancer treatment: An updated review. Genes Dis. 2018; 5:304–311. https://doi.org/10.1016/j.gendis.2018.06.003 PMID: 30591931

69. Archin NM, Kirchherr JL, Sung JAM, Clutton G, Sholtis K, Xu Y, et al. Interval dosing with the HDAC inhibitor vorinostat effectively reverses HIV latency. J Clin Invest. 2017; 127:3126–3135. https://doi.org/10.1172/JCI92684 PMID: 28714868

70. Gupta A, Chin W, Zhu L, Mok S, Luah Y, Lim E, et al. Dynamic epigenetic regulation of gene expression during the life cycle of malaria parasite *Plasmodium falciparum*. PLOS Pathogens. 2013; 9:e1003170. https://doi.org/10.1371/journal.ppat.1003170 PMID: 23468622